

## CLAIMS

- 1        1. A computer assisted method of auditing a superset of training data, the  
2        superset comprising examples of documents having one or more category  
3        assignments, the method including:
  - 4            partitioning the superset into at least two disjoint sets, including a test set and a  
5            training set, wherein the test set includes one or more test documents and the  
6            training set includes examples of documents belonging to at least two  
7            categories;
  - 8            categorizing the test documents using the training set;
  - 9            calculating a metric of confidence based on results of the categorizing step and  
10          the category assignments for the test documents; and
  - 11          reporting the test documents and category assignments that are suspicious and  
12          that appear to be missing, based on the metric of confidence.
- 1        2. The method of claim 1, further including repeating the partitioning,  
2        categorizing and calculating steps until at least one-half of the documents in the  
3        superset have been assigned to the test set.
- 1        3. The method of claim 2, wherein the test set created in the partition step has a  
2        single test document.
- 1        4. The method of claim 2, wherein the test set created in the partition step has a  
2        plurality of test documents.
- 1        5. The method of claim 1, further including repeating the partitioning,  
2        categorizing and calculating steps until substantially all of the documents in the  
3        superset have been assigned to the test set.
- 1        6. The method of claim 1, wherein the partitioning, categorizing and calculating  
2        steps are carried out substantially without user intervention.
- 1        7. The method of claim 5, wherein the partitioning, categorizing and calculating  
2        steps are carried out substantially without user intervention.
- 1        8. The method of claim 1, wherein the partitioning, categorizing, calculating and  
2        reporting steps are carried out substantially without user intervention.

1        9. The method of claim 5, wherein the partitioning, categorizing, calculating and  
2 reporting steps are carried out substantially without user intervention.

1        10. The method of claim 1, wherein the categorizing step includes determining k  
2 nearest neighbors of the test documents and the calculating step is based on a k  
3 nearest neighbors categorization logic.

1        11. The method of claim 10, wherein the metric of confidence is an unweighted  
2 measure of distance between the test document and the examples of documents  
3 belonging to various categories.

1        12. The method of claim 11, where the unweighted measure includes application  
2 of a relationship  $\Omega_0(d_t, T_m) = \sum_{d \in \{K(d_t) \cap T_m\}} s(d_t, d)$ , wherein

3         $\Omega_0$  is a function of the test document represented by the a feature vector  $d_t$  and of  
4 various categories  $T_m$ ; and

5         $s$  is a metric of distance between the test document feature vector  $d_t$  and certain  
6 sample documents represented by feature vectors  $d$ , the certain sample  
7 documents being among a set of k nearest neighbors of the test document having  
8 category assignments to the various categories  $T_m$ .

1        13. The method of claim 10, wherein the metric of confidence is a weighted  
2 measure of distance between the test document and the examples of documents  
3 belonging to various categories, the weighted measure taking into account the density  
4 of a neighborhood of the test document.

1        14. The method of claim 13 where the weighted measure includes application of

2        a relationship  $\Omega_1(d_t, T_m) = \frac{\sum_{d_1 \in \{K(d_t) \cap T_m\}} s(d_t, d_1)}{\sum_{d_2 \in K(d_t)} s(d_t, d_2)}$ , wherein

3         $\Omega_1$  is a function of the test document represented by the a feature vector  $d_t$  and of  
4 various categories  $T_m$ ; and

5         $s$  is a metric of distance between the test document feature vector  $d_t$  and certain  
6 sample documents represented by feature vectors  $d_1$  and  $d_2$ , the certain sample  
7 documents  $d_1$  being among a set of k nearest neighbors of the test document

8 having category assignments to the various categories  $T_m$  and the certain sample  
9 documents  $d_2$  being among a set of k nearest neighbors of the test document.

1 15. The method of claim 1, wherein the identifying step further includes filtering  
2 the test documents based on the metric of confidence.

1 16. The method of claim 15, wherein the filtering step further includes color  
2 coding the identified test documents based on the metric of confidence.

1 17. The method of claim 15, wherein the filtering step further includes selecting  
2 for display the identified test documents based on the metric of confidence.

1 18. The method of claim 1, wherein the user interface is a printed report.

1 19. The method of claim 1, wherein the user interface is a file conforming to  
2 XML syntax.

1 20. The method of claim 1, wherein the user interface is a sorted display  
2 identifying at least a portion of the test documents.

1 21. The method of claim 1, further including calculating a precision score for the  
2 identified test documents.

1 22. A computer assisted method of auditing a superset of training data, the  
2 superset comprising examples of documents having one or more category  
3 assignments, the method including:

4 determining k nearest neighbors of the documents in the superset;  
5 categorizing the documents based on the k nearest neighbors into a plurality of  
6 categories;  
7 calculating a metric of confidence based on results of the categorizing step and  
8 the category assignments for the documents; and  
9 reporting the documents and category assignments that are suspicious and that  
10 appear to be missing, based on the metric of confidence.

1 23. The method of claim 22, wherein the metric of confidence is an unweighted  
2 measure of distance between the test document and the examples of documents  
3 belonging to various categories.

1        24. The method of claim 23, where the unweighted measure includes application  
 2        of a relationship  $\Omega_0(\mathbf{d}_t, T_m) = \sum_{\mathbf{d} \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d})$ , wherein

3         $\Omega_0$  is a function of the test document represented by the a feature vector  $\mathbf{d}_t$  and of  
 4        various categories  $T_m$ ; and

5         $s$  is a metric of distance between the test document feature vector  $\mathbf{d}_t$  and certain  
 6        sample documents represented by feature vectors  $\mathbf{d}$ , the certain sample  
 7        documents being among a set of k nearest neighbors of the test document having  
 8        category assignments to the various categories  $T_m$ .

1        25. The method of claim 22, wherein the metric of confidence is a weighted  
 2        measure of distance between the test document and the examples of documents  
 3        belonging to various categories, the weighted measure taking into account the density  
 4        of a neighborhood of the test document.

1        26. The method of claim 25, wherein the weighted measure includes application

2        of a relationship  $\Omega_1(\mathbf{d}_t, T_m) = \frac{\sum_{\mathbf{d}_1 \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d}_1)}{\sum_{\mathbf{d}_2 \in K(\mathbf{d}_t)} s(\mathbf{d}_t, \mathbf{d}_2)}$ , wherein

3         $\Omega_1$  is a function of the test document represented by the a feature vector  $\mathbf{d}_t$  and of  
 4        various categories  $T_m$ ; and

5         $s$  is a metric of distance between the test document feature vector  $\mathbf{d}_t$  and certain  
 6        sample documents represented by feature vectors  $\mathbf{d}_1$  and  $\mathbf{d}_2$ , the certain sample  
 7        documents  $\mathbf{d}_1$  being among a set of k nearest neighbors of the test document  
 8        having category assignments to the various categories  $T_m$  and the certain sample  
 9        documents  $\mathbf{d}_2$  being among a set of k nearest neighbors of the test document.

1        27. The method of claim 22, wherein the determining, categorizing and  
 2        calculating steps are carried out substantially without user intervention.

1        28. The method of claim 22, wherein the identifying step further includes  
 2        filtering the documents based on the metric of confidence.

1        29. The method of claim 28, wherein the filtering step further includes color  
 2        coding the identified documents based on the metric of confidence.

- 1        30. The method of claim 28, wherein the filtering step further includes selecting  
2 for display the identified documents based on the metric of confidence.
- 1        31. The method of claim 22, wherein the user interface is a printed report.
- 1        32. The method of claim 22, wherein the user interface is a file conforming to  
2 XML syntax.
- 1        33. The method of claim 22, wherein the user interface is a sorted display  
2 identifying at least a portion of the documents.
- 1        34. The method of claim 22, further including calculating a precision score for  
2 the identified documents.